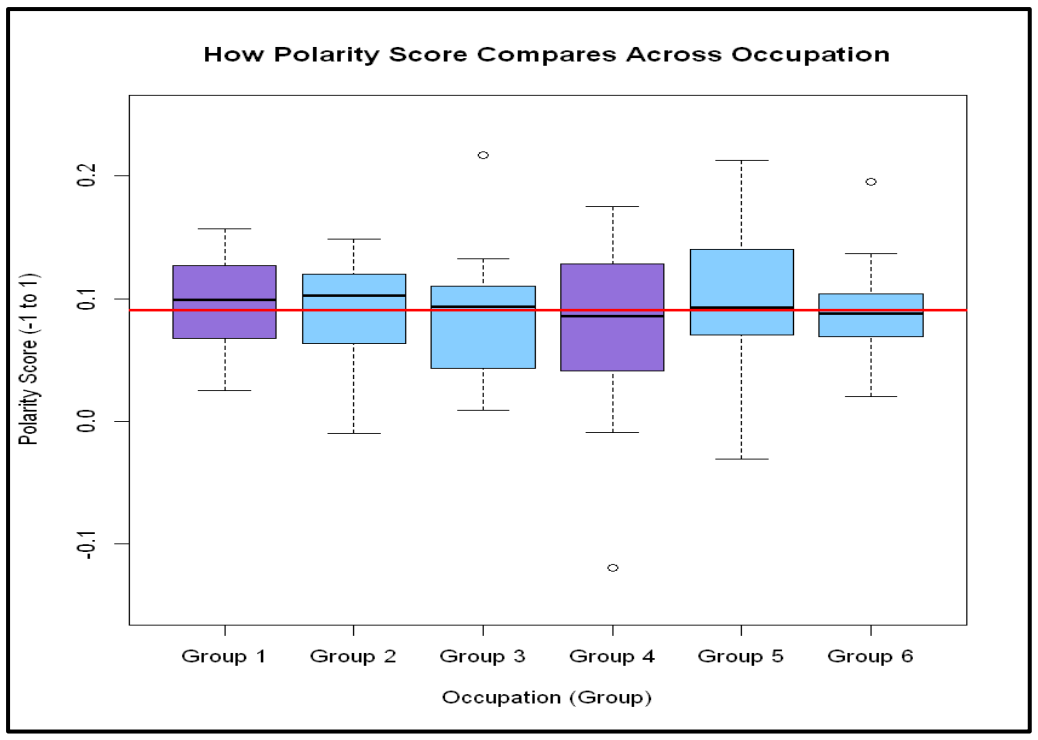
The University of Manchester

Q-Step | Data Fellow 2022

Natural Language Processing with Qualitative Data – UK Data Service

How did sentiments during the Foot and Mouth Disease outbreak compare across affected occupations?

Leila Peltier | **BA(Hons) Chinese and Linguistics**



Box plot showing polarity scores across Occupation

Overview of the Data Fellowship

This data fellowship involved using natural language processing (NLP) techniques to analyse the UKDS dataset on 'Health and Social Consequences of the Foot and Mouth Epidemic' (2001-2003). This project aimed to see how my background as a social science student would affect my experiences in using computational methods to do data analysis on qualitative data.

As part of this project, I used the dataset to research the sentiments of farmers and frontline workers (referred to as Groups 1 and 4) during the Foot and Mouth Epidemic in comparison to other affected occupations.

Data Analysis

My analysis involved comparing the average sentiment polarity score (from -1 to 1, where -1 is the most negative and 1 is the most positive) of Groups 1 and 4 to the polarity scores of other groups in the dataset using sentiment analysis.

Because the dataset is qualitative, I had to implement the text mining process before I could do any data visualisations or statistical testing. This process involved:

- Pre-processing data by converting rtf files into a Pandas DataFrame with each entry categorised by Occupation
- Processing by using NLP techniques such as word tokenisation and lemmatisation to standardise entries and make sentiment analysis more efficient
- Sentiment analysis of the processed data to get average polarity scores of each Occupation

I then used One-way ANOVA and Tukey's HSD test to create my findings.

Findings

The mean sentiment polarity score of the whole dataset was 0.0908 (4dp), which indicates more neutral sentiment. The mean polarity scores of Farmers and frontline workers were 0.0974 (4dp) and 0.0772 (4dp) respectively, which is similar to the overall polarity score.

No significant difference in polarity score across occupations was found $F(5, 81) = 0.384$, $p = .859$. Farmers and frontline workers also did not have significantly different polarity scores compared to the other occupations.

Key Skills Learnt

In terms of technical skills, I have learned how to:

- Use Python to implement data science techniques such as cleaning and manipulating data
- Use Python to implement Natural Language Processing techniques such as sentiment analysis, stemming, and lemmatisation
- Use R to visualise data and conduct statistical tests
- Use Jupyter Notebooks and GitHub to conduct collaborative projects
- Use Microsoft Teams to implement schedules, delegate and prioritise tasks in a project, set goals, conduct meetings, and share files with co-workers

In learning these technical skills, I also developed key analytical and research skills. These include:

- Formulating an appropriate research question
- Identifying areas of inconsistent formatting and gaps in a dataset
- Synthesising information using qualitative and quantitative research methods
- Critically assessing findings to draw conclusions on my research

Through this data fellowship, I also learned professional skills such as:

- Communication, collaboration and teamwork when conducting projects
- Time management and organisation
- Problem-solving and adaptability in tackling wide ranges of data processing and coding challenges

	Occupation	Polarity_Score_Mean
1	Group 1	0.09741172
2	Group 2	0.09077748
3	Group 3	0.08688578
4	Group 4	0.07716527
5	Group 5	0.10130420
6	Group 6	0.09098647

Polarity mean score by Occupation

Contact: : **l_peltier@outlook.com**